基于 LDA 模型的文本分割

石 晶[1] 胡 明[1] 石 鑫[2] 戴国忠[3]
[1](长春工业大学计算机科学与工程学院 长春 130012)
[2](辽宁师范大学功能材料化学研究所 辽宁 大连 116029)
[3](中国科学院软件研究所人机交互技术与智能信息处理实验室 北京 100190)

摘 要 文本分割在信息提取、文摘自动生成、语言建模、首语消解等诸多领域都有极为重要的应用. 基于 LDA 模型的文本分割以 LDA 为语料库及文本建模, 利用 MCMC 中的 Gibbs 抽样进行推理, 间接计算模型参数, 获取词汇的概率分布, 使隐藏于片段内的不同主题与文本表面的字词建立联系. 实验以汉语的整句作为基本块, 尝试多种相似性度量手段及边界估计策略, 其最佳结果表明二者的恰当结合可以使片段边界的识别错误率远远低于其它同类算法.

关键词 文本分割; LDA 模型; 相似性度量; 边界识别
中图法分类号 TP301

Text Segmentation Based on Model LDA

SHI Jing[1] HU Ming[1] SHI Xin[2] DAI Guo-Zhong[3]
[1](*School of Computer Science and Engineering, Changchun University of Technology, Changchun* 130012)
[2](*Institute of Chemistry for Functionalized Materials, Liaoning Normal University, Dalian, Liaoning* 116029)
[3](*Computer Human Interaction and Intelligent Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing* 100190)

Abstract Text segmentation is very important for many fields including information retrieval, summarization, language modeling, anaphora resolution and so on. Text segmentation based on LDA models corpora and texts with LDA. Parameters are estimated with Gibbs sampling of MCMC and the word probability is represented. Different latent topics are associated with observable words. In the experiments, Chinese whole sentences are taken as elementary blocks. Variety of similarity metrics and several approaches of discovering boundaries are tried. The best results show the right combination of them can make the error rate far lower than other algorithms of text segmentation.

Keywords text segmentation; model Latent Dirichlet Allocation (LDA); similarity metric; boundaries discovering

1 引 言

文本分割是指在一个书面文档或语音序列中自动识别具有独立意义的单元(片段)之间的边界, 其分割对象可以是语音流、网络动态数据, 或者书面静态文本. 这种预处理在很多领域都有极为重要的应用, 比如信息提取、文摘自动生成、文本解析、语言建

收稿日期: 2006-10-09; 最终修改稿收到日期: 2007-11-27. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2002CB312103)、国家自然科学基金(60503054)和中国科学院软件研究所创新工程重大项目资助. 石 晶, 女, 1970 年生, 博士, 主要从事自然语言理解及信息处理方面的研究. E-mail: crystal1087@126.com. 胡 明, 1963 年生, 博士, 主要从事人工智能方面的研究. 石 鑫, 1975 年生, 副教授, 主要从事计算机应用方面的研究. 戴国忠, 男, 1944 年生, 研究员, 总工程师, 从事软件工程和计算机图形学工程的研究, 致力于图形标准化软件的研制和用户界面的开发.

模、文本导航以及首语消解等.

目前常用的分割方法大致基于如下几种思想:(1)假定相同、相似或语义相关的词汇倾向于出现在同一片段内[1-2];(2)认为特定的语言现象,比如提示短语、停顿标记、韵律特征、指代、句法及词汇的形态同化等与片段首尾隐含某种必然联系[3-4];(3)相信合适的概率统计模型能够为片段边界的估计提供可靠依据[5-6].对于第三种方法,选择合适的模型是保证分割效果的关键.文献[6]采用 PLSA 模型,但模型中的文档概率值与特定文档相关,因此缺乏处理新文档的自然方法.同时待估参数的数量随着文档数量的增多线性增长,说明模型易于过度拟合.与 PLSA 模型相比,LDA(Latent Dirichlet Allocation)[7]称得上是完全的生成模型.由于该模型将主题混合权重视为 k 维参数的潜在随机变量,而非与训练数据直接联系的个体参数集合,推理上采用 Laplace 近似、变分近似、MCMC(Markov chain Monte Carlo)[8]以及期望-扩散(expectation-propagation)[9]等方法获取待估参数值,所以克服了 PLSA 的不足.本文介绍的分割方法即以 LDA 为语料库及文本建模,利用 MCMC 中的 Gibbs 抽样近似估算模型参数,获取词汇的概率分布.以汉语的整句作为分割的基本块,除了尝试多种相似性度量手段,还尝试不同的边界识别策略.实验表明,最佳结果(不考虑常数法识别边界的情况)的错误率远远低于其它同类算法.

近几年,LDA 模型、LDA 的扩展模型以及它们在自然语言和智能信息处理中的应用得到充分的重视和深入的研究,扩展模型包括空间 LDA 模型[10]、作者-角色-主题模型[11]等;应用涉及词义排歧[12]、词性标注[13]、主题分解[14]、信息抽取[15]等,但还没有人基于 LDA 实现文本的主题分割,本文恰恰进行了这一方面的尝试.

本文第2节介绍 LDA 模型;第3节详述本文的分割策略;第4节给出测试手段及实验结果,并就实验结果进行讨论;第5节对比、分析最新相关研究及工作;最后总结全文.

2 LDA 模型

目前的概率主题模型一般基于同样的思想——文本是若干主题的随机混合.不同的模型会进一步作不同的统计假设,以不同的方式获取模型参数.

2.1 模型介绍

一个文本通常需要讨论若干主题,而文本中的特定词汇体现出所讨论的特定主题.在统计自然语言处理中,为文本主题建模的方法是视主题为词汇的概率分布,文本为这些主题的随机混合.假设有 T 个主题,则所给文本中的第 i 个词汇 w_i 可以表示如下:

$$P(w_i)=\sum_{j=1}^{T}P(w_i|z_i=j)P(z_i=j) \qquad (1)$$

其中,z_i 是潜在变量,表明第 i 个词汇记号 w_i 取自该主题,$P(w_i|z_i=j)$ 是词汇 w_i 记号属于主题 j 的概率,$P(z_i=j)$ 给出主题 j 属于当前文本的概率.假定 T 个主题形成 D 个文本以 W 个唯一性词汇表示,为记号方便,令 $\varphi_w^{(z=j)}=P(w|z=j)$ 表示对于主题 j,W 个词汇上的多项分布,其中 w 是 W 个唯一性词汇表中的词汇;令 $\psi_{=j}^{(d)}=P(z=j)$ 表示对于文本 d,T 个主题上的多项分布,于是文本 d 中词汇 w 的概率为

$$P(w|d)=\sum_{j=1}^{T}\varphi_w^{(z=j)}\cdot\psi_{=j}^{(d)} \qquad (2)$$

LDA 模型[6]在 $\psi^{(d)}$ 上作 $Dirichlet(\alpha)$ 的先验概率假设,使得模型易于处理训练语料之外的新文本.为了便于模型参数的推理,本文除了在 $\psi^{(d)}$ 上作对称的 $Dirichlet(\alpha)$ 的先验概率假设外,在 $\varphi^{(z)}$ 上亦作对称的 $Dirichlet(\chi)$ 的先验概率假设[9],如下:

$$w_i|z_i,\varphi^{(z_i)}\sim Discrete(\varphi^{(z_i)}),\quad \varphi^{(z_i)}\sim Dirichlet(\chi),$$
$$z_i|\psi^{(d_i)}\sim Discrete(\psi^{(d_i)}),\qquad \psi^{(d_i)}\sim Dirichlet(\alpha),$$

这里的 χ 可以理解为,在见到语料库的任何词汇之前,从主题抽样获得的词汇出现频数,而 α 可以理解为,在见到任何文档文字之前,主题被抽样的频数.尽管 α 和 χ 的具体取值会影响到主题及词汇被利用的程度,但不同的主题被利用的方式几乎没有变化,不同的词汇被利用的方式也基本相同,因此可以假定对称的 dirichlet 分布,即所有的 α 取相同的值,所有的 χ 取相同的值.

2.2 Gibbs 抽样

为了获取词汇的概率分布,本文没有将 φ 和 ψ 作为参数直接计算,而是考虑词汇对于主题的后验概率 $P(w|z)$,利用 Gibbs 抽样间接求得 φ 和 ψ 的值.MCMC 系一套从复杂的概率分布抽取样本值的近似迭代方法,Gibbs 抽样作为 MCMC 的一种简单实现形式,其目的是构造收敛于某目标概率分布的 Markov 链,并从链中抽取被认为接近该概率分布值的样本.于是目标概率分布函数的给出便成为使用 Gibbs 抽样的关键.对于本文的 LDA 模型,仅仅需要对主题的词汇分配,也就是变量 z_i 进行抽样.记后验概率为 $P(z_i=j|z_{-i},w_i)$,计算公式如下(详见

附录):

$$P(z_i=j\mid z_{-i},w_i)=\frac{\dfrac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W\chi}\cdot\dfrac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}}{\sum_{j=1}^{T}\dfrac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W\chi}\cdot\dfrac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}} \tag{3}$$

其中, $z_i=j$ 表示将词汇记号 w_i 分配给主题 j, 这里 w_i 被称为词汇记号是因为其不仅代表词汇 w, 而且与该词所在的文本位置相关, z_{-i} 表示所有 z_k $(k\neq i)$ 的分配. $n_{-i,j}^{(w_i)}$ 是分配给主题 j 与 w_i 相同的词汇个数; $n_{-i,j}^{(\cdot)}$ 是分配给主题 j 的所有词汇个数; $n_{-i,j}^{(d_i)}$ 是文本 d_i 中分配给主题 j 的词汇个数; $n_{-i,\cdot}^{(d_i)}$ 是 d_i 中所有被分配了主题的词汇个数; 所有的词汇个数均不包括这次 $z_i=j$ 的分配.

Gibbs 抽样算法详述如下:

1. z_i 被初始化为 1 到 T 之间的某个随机整数. i 从 1 循环到 N, N 是语料库中所有出现于文本中的词汇记号个数. 此为 Markov 链的初始状态.

2. i 从 1 循环到 N, 根据式(3)将词汇分配给主题, 获取 Markov 链的下一个状态.

3. 迭代第 2 步足够次数以后, 认为 Markov 链接近目标分布, 遂取 z_i (i 从 1 循环到 N)的当前值作为样本记录下来. 为了保证自相关较小, 每迭代一定次数, 记录其它的样本.

舍弃词汇记号, 以 w 表示唯一性词, 对于每一个单一样本, 可以按下式估算 φ 和 ψ 的值:

$$\varphi_w^{(z=j)}=\frac{n_j^{(w)}+\chi}{n_j^{(\cdot)}+W\chi},\qquad \psi_{z=j}^{(d)}=\frac{n_j^{(d)}+\alpha}{n_\cdot^{(d)}+T\alpha} \tag{4}$$

其中, $n_j^{(w)}$ 表示词汇 w 被分配给主题 j 的频数; $n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数; $n_j^{(d)}$ 表示文本 d 中分配给主题 j 的词数; $n_\cdot^{(d)}$ 表示文本 d 所有被分配了主题的词数.

3　分割策略

待分割文本是语料库训练时没有处理过的新文本, 如果对于每一个未知文本, 都将其加入语料库后重新训练, 则异常浪费时间, 亦没有必要, 本文的做法是只对新加入的词汇记号运行 Gibbs 抽样算法, 且只迭代较少的次数. 预处理的基本块采用汉语的整句 s, 分割的大致步骤如下:

1. 对于语料库文本的词汇记号运行 Gibbs 抽样算法, 迭代足够次;

2. 以整句 s 作为式(3)中的文本 d, 遍历待分割文本的所有词汇记号, 运行 Gibbs 抽样算法, 迭代少数几次;

3. 按照式(4)分别计算 φ 和 ψ 的值;

4. 根据公式 $P(w\mid s)=\sum_{j=1}^{T}\varphi_w^{(z=j)}\cdot\psi_{z=j}^{(s)}$ 求取待分割文本词汇的概率分布 $P(w\mid s)$;

5. 基于 $P(w\mid s)$, 利用不同的度量手段计算句间的相似值 Sim;

6. 结合局部最小值的边界估计策略, 通过句间相似值 Sim 识别片段边界.

3.1　相似性度量

基于 $P(w\mid s)$ 计算句间的相似值, 需要选择合适的度量手段, 本文尝试如下 5 种方法:

(1)余弦度量

$$Sim_{\mathrm{Cos}}=\frac{\sum_{w\in W}P(w\mid s_1)P(w\mid s_2)}{\sqrt{\sum_{w\in W}P(w\mid s_1)^2}\sqrt{\sum_{w\in W}P(w\mid s_2)^2}} \tag{5}$$

(2)L1 距离度量[16]

$$Sim_{\mathrm{L1}}=1-\frac{\sum_{w\in W}|P(w\mid s_1)-P(w\mid s_2)|}{2} \tag{6}$$

(3)Hellinger 距离度量

$$Sim_{\mathrm{Hel}}=\sum_{w\in W}\sqrt{P(w\mid s_1)P(w\mid s_2)} \tag{7}$$

(4)Clarity 度量[17]

$$Sim_{\mathrm{Clr}}=-KL(P(w\mid s_1)\|P(w\mid s_2))+KL(P(w\mid s_1)\|GC)-KL(P(w\mid s_2)\|P(w\mid s_1))+KL(P(w\mid s_2)\|GC) \tag{8}$$

其中, GC 代表词汇 w 在训练语料库的出现频率, 即 $f(w)$, $KL(\cdot\|\cdot)$ 被称为相对熵:

$$KL(P(w\mid s_1)\|P(w\mid s_2))=\sum_{w\in W}P(w\mid s_1)\log_2\frac{P(w\mid s_1)}{P(w\mid s_2)} \tag{9}$$

(5)Jensen-Shannon 发散度量

$$Sim_{\mathrm{JS}}=-KL\left(P(w\mid s_1)\left\|\frac{P(w\mid s_1)+P(w\mid s_2)}{2}\right.\right)-KL\left(P(w\mid s_2)\left\|\frac{P(w\mid s_1)+P(w\mid s_2)}{2}\right.\right) \tag{10}$$

3.2　边界识别

利用不同的边界估计策略进行文本分割的结果显见不会相同, 本文对比如下 4 种方法, 以求探究最佳策略:

(1)阈值法

设定常数 θ, 若句间相似值 $Sim(s_1,s_2)<\theta$, 则认为 s_1, s_2 分属于不同的片段. 该方法极易实现, 如果所给 θ 合适, 可以获得较低的错误率.

(2)动态常数法

阈值法虽然简单, 但需要人为设定 θ, 很难给出最佳值, 因此可以考虑根据相邻句间的相似值表动态

改变 θ. 假设待分割文本有 n 个整句, 则相邻句间的相似值表为 $SimTable=\{Sim_1, Sim_2, \cdots, Sim_i, \cdots, Sim_{n-1}\}$, 其中 $Sim_i=Sim(s_i, s_{i+1})$, $1\leq i\leq n-1$, 令

$$avgSim=\frac{Sim_1+Sim_2+\cdots+Sim_i+\cdots+Sim_{n-1}}{n-1},\quad 1\leq i\leq n-1,$$

$$avgmSim=\frac{(Sim_2-Sim_1)+\cdots+(Sim_{n-1}-Sim_{n-2})}{n-2},$$

若 $avgmSim\leq Sim(s_1, s_2)\leq avgSim$, 则认为 s_1, s_2 分属于不同的片段.

(3) 局部最小值法[6]

在相邻句间的相似值表 $SimTable$ 中选择局部最小值 $Sim_{min}(s_1, s_2)$; 从每一个局部最小值出发向左、向右分别寻找距离最近的较大值 Sim_{maxl} 以及 Sim_{maxr}, 利用公式 $d_{rel}(s_1, s_2)=\frac{Sim_{maxl}+Sim_{maxr}}{2Sim_{min}(s_1, s_2)}-1$ 计算相对深度; 令 α 为一常数, 若相对深度 $d_{rel}(s_1, s_2)\geqslant\alpha$, 则 s_1, s_2 分属于不同的片段.

(4) 动态规划法[18]

将某文本分割为 K 个片段 $\{t_1, t_2, \cdots, t_K\}$, 片段 $t_k(1\leq k\leq K)$ 由首尾句子 s_i, s_j 决定, 或者 $t_k=[i, j]$, 于是 t_k 的平均相似值为

$$\beta(t_k)=\frac{\sum_{i\in t_k}\sum_{j\in t_k}Sim(s_i, s_j)}{(j-i+1)^2}.$$

类似于文献[15] 定义如下分割代价:

$$J(t;\mu,\sigma,r,\gamma)=\sum_{k=1}^{K}\left(\gamma\frac{(t_k-t_{k-1}-\mu)}{2\sigma^2}-(1-\gamma)\frac{\sum_{t=t_{k-1}+1}^{t_k}\sum_{t=t_{k-1}+1}^{t_k}\beta(t_k)}{(t_k-t_{k-1})^r}\right)\tag{11}$$

其中, μ 和 σ 是片段长度的数学期望及均方差, 可以通过训练语料库统计获得, r 和 γ 试探给值, 本实验取 $\gamma=0.9$, $r=0.3333$. 利用动态规划算法求解使得代价式(14)取最小值的分割.

4 实验结果及讨论

本文所有实验以 1998 年《人民日报》手工标注的语料库为背景库及建模对象(共 3157 个文本), 并以知网词典(去除其中的虚词、形容词、副词等意义不大的词, 再删掉语料库出现频数小于 5 的词, 剩余 18049 个词汇)作为选择词汇的词典.

为了有效利用 Gibbs 抽样法, 先通过实验确定主题数目 T 的最佳值以及 burn-in 间距和 thinning 间距的取值, 然后对文本分割进行测试.

4.1 主题数目的确定

针对同样的语料库及同样的词典($W=18049$, $D=3157$, $N=562931$, W 为词汇数目, D 为文本数目, N 为词汇记号数目, 也就是每次抽样依据式(3)对 z 赋值的次数), 可变量包括超参数 α, χ以及主题数目 T. 本实验目的在于了解主题数目对于 Gibbs 抽样算法的影响, 为此先确定 α, χ的值, 然后为 T 选择合适的值. 这实际上是一个模型选择的问题, 本文采用贝叶斯统计中的标准方法予以解决. 令 $\alpha=50/T$, $\chi=0.01$(此为经验值, 多次实验表明, 这种取值在本实验的语料库上有较好表现), T 取不同的值分别运行 Gibbs 抽样算法, 检测 $\log P(w|T)$ 值的变化.

由本文建模的模型知, α, χ是多项分布 ψ和 φ上的 Dirichlet 先验概率假设, 其自然共轭的特点说明通过对 ψ和 φ积分可以求取联合概率 $P(w, z)$ 的值. $P(w, z)=P(w|z)P(z)$, 并且 φ和 ψ分别单独出现于第 1 项和第 2 项, 对 φ积分获第 1 项值如下:

$$P(w|z)=\left(\frac{\Gamma(W\chi)}{\Gamma(\chi)^W}\right)^T\prod_{j=1}^{T}\frac{\prod_w\Gamma(n_j^{(w)}+\chi)}{\Gamma(n_j^{(\cdot)}+W\chi)}\tag{12}$$

其中, $\Gamma(\cdot)$ 是标准的 gamma 函数, $n_j^{(w)}$ 表示词汇 w 分配给主题 j 的频数, $n_j^{(\cdot)}$ 表示分配给主题 j 的所有词数. 因为 $P(w|T)$ 可以近似为一系列 $P(w|z)$ 的调和平均值, 所以按下式求取其值:

$$\frac{1}{P(w|T)}=\frac{1}{M}\sum_{m=1}^{M}\frac{1}{P(w|z^{(m)})}\tag{13}$$

实验结果如图 1 所示.

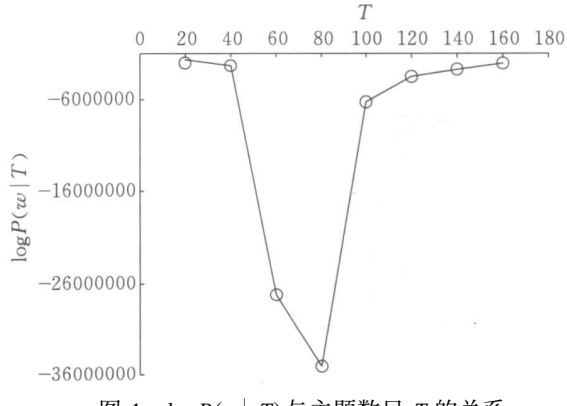

图 1 $\log P(w|T)$ 与主题数目 T 的关系

由图 1 可以看出, 当主题数目 T 为 80 时, $\log P(w|T)$ 的值最小, 随后开始急遽增大, 说明主题数目为 80 时, 模型对于语料库数据中有效信息的拟合最佳, 因此, 后续实验的主题数目取为 80.

4.2　Burn-in 及 Thinning 间距的选择

Gibbs 抽样算法从初始值开始运行, 迭代足够次 b 后认为样本接近目标概率分布, 然后每隔一定次数 c 抽取样本, b 称为 burn-in 间距, c 称为 thinning 间距. b 和 c 的取值比较难以确定, 一般与特定的语料库相关. 如果所构造 Markov 链的相邻状态间关联较小, b, c 以较小的值可满足需要, 但如果相邻状态间的关联较大, 就必须增大 b, c 的取值, 方可降低自相关.

本实验取 $T=80$, 以 4 次不同的初始值运行 Gibbs 算法, 若 b, c 的取值合适, 则抽样结果($\log P(w|z)$的值)随初始值的变化很小, 也可以说独立于初始值. 实验结果如图 2 所示.

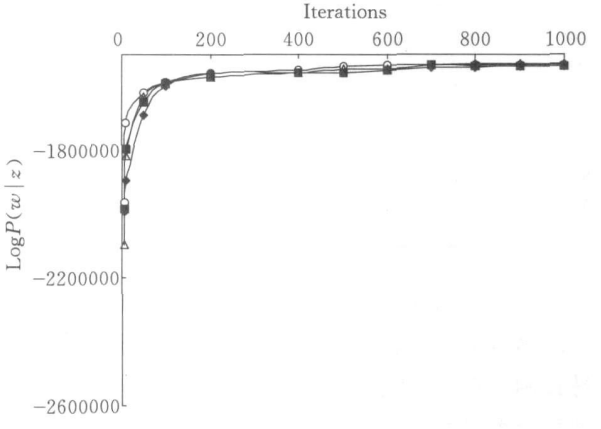

图 2　迭代数百次后 $\log P(w|z)$趋于稳定

从图 2 中可以看出, $\log P(w|z)$的值在迭代数百次后稳定, 因此本文实验取 burn-in 间距为 1000, thinning 间距为 100.

4.3　测试集及度量标准

文本分割算法的评测标准比较主观, 人们对于片段边界的位置以及文本分割的粒度往往没有一致的看法和观点, 这就为分割结果的判断增加了很大的难度. 为了解决这个问题, 一部分研究将不同内容的文本连接起来, 人为决定片段边界. 另外一部分研究则按人的判断估价, 采用大多数人的意见作为标准. 为了获得客观的评测结果, 本文采用第一种策略.

本实验利用 1997 年 3 月份《人民日报》手工标注的语料库构建 4 个测试集 T_{3-11}, T_{3-5}, T_{6-8}, T_{9-11}, T_{x-y}表示所含主题片段的句数在 x 和 y 之间. 每一个测试集包括若干伪文本, 即由不同类的文本连接而成的形式上的文本, 要求相邻段落务必来自不同的类. 其所含的主题数平均为 7, 具体如表 1.

表 1　实验中的测试集

	句数①	伪文本数
T_{3-11}	3－11	109
T_{3-5}	3－5	127
T_{6-8}	6－8	115
T_{9-11}	9－11	98

注: ① 为伪文本中每个主题片段的句数.

为了便于同类算法的对比, 本文采取两种度量标准, 错误率 P_k[19] 和 WindowDiff[20].

$$P_k=P(\text{seg})P(\text{miss})+(1-P(\text{seg}))P(\text{false alarm}) \tag{14}$$

$P(\text{seg})$是指距离为 k 的两个句子分属不同主题片段的概率, 而 $1-P(\text{seg})$就是指距离为 k 的两个句子属于同一主题片段的概率, 本实验将两个先验概率取等值, 即 $P(\text{seg})=0.5$, $P(\text{miss})$是算法分割结果缺少一个片段的概率, $P(\text{false alarm})$是算法分割结果添加一个片段的概率.

$$WindowDiff(ref,hyp)=\frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i,ref_{i+k})-b(hyp_i,hyp_{i+k})|>0) \tag{15}$$

其中, $b(i,j)$表示整句 s_i和整句s_j间的边界数量, N 表示文本中的整句数量, k 取真实片段平均长度的一半, ref 代表真实分割, hyp 代表算法分割.

4.4　实验结果及讨论

首先在表 2 列出实验叙述中所涉及的记号含义.

表 2　记号含义

记号	含义
Cos	Cosine distance
Hel	Hellinger distance
Clr	Clarity metric
JS	Jensen-Shannon divergence
L1	L1 distance
Con	Constant
DyCon	Dynamic constant
Loc	Local minimum
DyPro	Dynamic programming
WD	WindowDiff
P_k	Error rate

4.4.1　不同测试集上的实验结果

Gibbs 抽样的主题数目 $T=80$, 超参数 $\alpha=50/T$, $\chi=0.01$. 取 10 个不同的初始值运行算法, 每个初始值迭代 1000 次, 然后每隔 100 次取一次样本, 共取 10 次样本. 加入训练语料的测试文本被初始化, 继续迭代 10 次, 开始计算结果. 每个文本的测试结果取 100 个样本的平均值, 测试集的实验结果取所有

文本测试结果的平均值. 其实验结果如表 3, 其中 $Mod_{x,y}$ 表示相似性度量采用 x 方式, 边界识别采用 y 策略, 二者的不同组合形成 4×5=20 种模型. 表中阈值法的常数值依次取 0.018, 0.29, 2.46, 0.099, 1.24.

表 3 不同测试集的错误率

Models	T_{3-11}		T_{3-5}		T_{6-8}		T_{9-11}	
	P_k/%	WD/%	P_k/%	WD/%	P_k/%	WD/%	P_k/%	WD/%
$Mod_{Cos,Con}$	7.14	18.57	7.94	15.87	5.26	11.05	7.29	18.18
$Mod_{Cos,DyCon}$	6.00	24.00	8.73	17.46	9.21	23.96	6.37	29.09
$Mod_{Cos,Loc}$	11.43	20.28	10.31	20.63	7.55	28.30	10.91	25.45
$Mod_{Cos,DyPro}$	10.00	24.28	9.52	19.04	8.49	24.53	10.93	30.91
$Mod_{Hel,Con}$	7.04	15.52	9.52	19.05	6.60	24.53	5.45	21.82
$Mod_{Hel,DyCon}$	12.86	21.73	9.40	29.70	7.46	17.77	13.28	32.73
$Mod_{Hel,Loc}$	11.48	25.46	8.83	18.41	9.94	19.60	9.09	21.81
$Mod_{Hel,DyPro}$	7.54	22.36	11.90	23.80	11.38	26.81	5.45	29.09
$Mod_{Cls,Con}$	9.74	28.36	10.31	20.63	7.30	27.68	4.64	14.55
$Mod_{Cls,DyCon}$	8.66	16.14	9.38	37.50	11.12	20.92	15.63	31.25
$Mod_{Cls,Loc}$	7.55	16.12	12.29	27.03	5.72	18.18	10.67	24.32
$Mod_{Cls,DyPro}$	11.47	32.78	8.90	25.6	10.80	26.60	12.73	34.55
$Mod_{JS,Con}$	8.44	28.75	6.89	13.79	9.90	22.00	12.50	31.25
$Mod_{JS,DyCon}$	7.50	29.37	9.90	21.60	8.14	27.36	9.38	25.00
$Mod_{JS,Loc}$	11.88	21.11	10.31	24.14	8.90	22.00	6.36	20.00
$Mod_{JS,DyPro}$	9.62	19.87	11.90	23.81	9.90	16.60	6.37	29.09
$Mod_{L1,Con}$	8.07	16.14	9.64	22.22	10.18	29.92	7.28	14.55
$Mod_{L1,DyCon}$	7.45	14.91	11.93	24.62	12.00	31.52	8.18	25.45
$Mod_{L1,Loc}$	10.24	20.49	10.32	20.63	7.26	26.72	14.55	36.36
$Mod_{L1,DyPro}$	11.80	23.60	15.62	33.75	8.74	20.18	14.06	28.13

从表 3 数据可以看出, 不同模型的错误率相对较接近, 基本集中在 5%~12%之间. 为了更清楚不同相似性度量及边界识别策略在 4 种测试集上的表现, 将其关系分别汇于图 3、图 4(错误率取最小值).

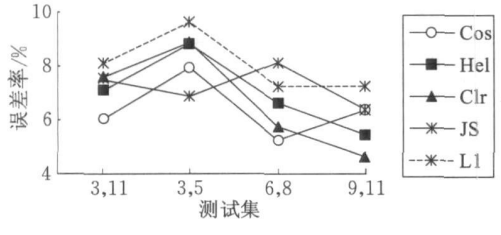

图 3 不同相似性度量手段在 T_{3-11}, T_{3-5}, T_{6-8}, T_{9-11} 的错误率(P_k)

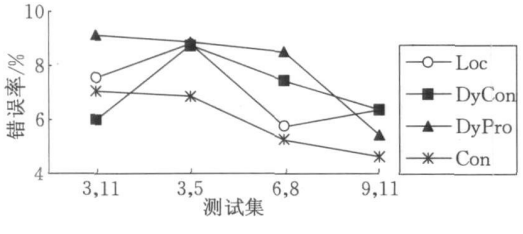

图 4 不同边界识别策略在 T_{3-11}, T_{3-5}, T_{6-8}, T_{9-11} 的错误率(P_k)

对于 T_{3-11}和 T_{6-8}, Cos 度量可以取得最小值, JS 度量适合 T_{3-5}, 而 Clr 度量在 T_{9-11} 上有好的表现.

由图中可见, 常数法可以为所有的测试集提供令人满意的边界识别, 但该方法有太大的随机性, 不易控制. 剔除常数法不考虑, 动态常数法在 T_{3-11} 上的错误率最低; T_{3-5} 由于片段内句数少, 反映主题的信息少, 所以没有特别好的边界估计策略; 局部最小值法适合 T_{6-8}; 而动态规划法在 T_{9-11} 上有最佳表现.

4.4.2 与其它分割算法的比较

作为与本文方法的对比, 取 PLSA[6]、LSA[7]、MDA[21] 3 种算法在 T_{3-11}, T_{3-5}, T_{6-8}, T_{9-11} 上进行测试, 错误率 P_k 如表 4.

表 4 与 PLSA、LSA 以及 MDA 的对比结果

	P_K/%			
	T_{3-11}	T_{3-5}	T_{6-8}	T_{9-11}
本文算法①	6.00	8.73	5.72	5.45
PLSA 算法	16.79	13.81	13.26	11.94
LSA 算法	13.12	15.21	10.02	12.17
MDA(多元判别)	11.61	11.38	11.94	11.00

注: ①取常数法之外的最佳结果.

可见, 基于 LDA 模型, 可以使分割的错误率远远低于其它模型及方法, 而且, 实验表明测试结果比较稳定, 不同样本间的差别较小. 本文作者曾对基于 PLSA 模型的文本分割进行仔细研究[22], 发现基于 PLSA 模型的分割, 其结果的随机性较大, 随迭代次数及主题数目的变化难以确定.

5　相关研究对比

本文探讨适合文本主题分割的模型和方法, 与本文研究相关的近期工作包括 STM[11] 、MDA[21] 以及 FBS[23] 等.

STM(Statistical Topic Model)是一种有限混合模型, 原则上讲, 这种模型假定一个文档仅仅呈现一个主题, 往往无法准确描述语料库及文档建模所需的数据信息, 同时, 由于没有对主题概率及词汇概率作任何假设, 导致局部极大值、过度拟合以及收敛速度过慢等问题. 本文作者在实验中发现, 基于该模型的主题分割错误率较高, 基本在 50%左右. 分析原因, 除了上述模型自身存在的问题外, 还由于对模型参数的估算基于单一文档的部分信息(包括 h 句的块), 而非语料库丰富的知识, 但毕竟一个块内提供的信息过于有限, 所以无法准确估算参数值.

MDA (Multiple Discriminant Analysis)方法定义了 4 种全局评价函数, 寻找满足分割单元内距离最小化和分割单元间距离最大化条件的最好分割方式, 实现对文本分割模式的全局评价. 其优点在于通用性强, 无需语料库, 缺点是片段边界的确定仅仅依赖本文档的内部信息, 难以实现更好的分割. 表 4 的实验结果同样说明, 采用 MDA 方法, 其分割错误率(P_k)极为集中(11%左右). 而本文方法更多地依赖于语料库的训练, 因此当语料库信息充分, 测试文档与训练语料结构类似时就会呈现更好的分割效果(P_k=5.72%).

FBS (Feature-Based Segmentation)方法首先选择合适的特征, 然后基于某种学习策略, 将文本分割看作分类问题予以解决, 其思想具有独特性, 但该方法的分割结果完全依赖于所选特征, 而特征的选择与确定是一个有待解决的问题, 尤其对于汉语. 同时该方法基于有效命名实体的判别, 且需要诸如 wordnet 的同义词辞典, 而汉语的类似资源极为匮乏, 尝试基于 FBS 的汉语文本分割目前还有困难.

6　结　语

基本块设置、相似性度量以及边界识别是文本分割系统的 3 个组成部分. 本文以 LDA 为语料库及文本建模, 将汉语的整句作为基本块, 尝试不同相似性度量手段与边界估计策略. 由于 LDA 是完全的生成模型, 从理论上讲, 具有其它模型无可比拟的建模优点. 实验结果表明, 基于 LDA 模型的文本分割无论采用何种相似性度量及边界识别方法, 均获较低的错误率, 基本集中在 5%～12%之间, 确有很好的分割效果.

文本分割除了需要直接测试, 如本文实验所示, 更需要间接测试, 即将其置入应用系统中考查. 本文研究的目的是为文本推理提供预处理, 所以下一步的工作将是对该方法更有效的测试.

致　谢　本文在研究中用到一些基础性的工作, 包括电子常识知识库《知网》, 汉语词法分析系统 ICTCLAS, 《人民日报》手工标注语料库以及文本分类语料库. 所有这些资源可以网上下载, 限于研究使用. 基于这样一些宝贵的资源, 我的研究得以进行与开展, 因此在这里对开发、设计、整理者表示由衷的感谢!

参　考　文　献

[1] Bolshakov Igor A, Gelbukh A. Text segmentation into paragraphs based on local text cohesion//Václav Matousek, Pavel Mautner, Roman Moucek, Karel Tauser eds. Proceedings of the Text, Speech and Dialogue (TSD-2001). Lecture Notes in Artificial Intelligence, N 2166. Springer-Verlag, 2001: 158-166

[2] Kehagias Ath, Nicolaou A, Fragkou P, Petridis V. Text segmentation by product partition models and dynamic programming. Mathematical and Computer Modelling, 2004, 39: 209-217

[3] Tur G, Hakkani-Tur D, Stolcke A, Shriberg E. Integrating prosodic and lexical cues for automatic topic segmentation. Computational Linguistics, 2001, 27(1): 31-57

[4] Levow Gina-Anne. Prosody-based topic segmentation for mandarin broadcast news//Proceedings of the HLT-NAACL 2004. Boston, Massachusetts, USA, 2004, 2: 137-140

[5] Blei D, Moreno P. Topic segmentation with an aspect hidden Markov model//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Louisiana, USA, 2001: 343-348

[6] Thorsten Brants, Francine Chen, Ioannis Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis//Proceedings of the 11th International Conference on Information and Knowledge Management McLean. Virginia, USA, 2002: 211-218

[7] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, (3): 993-1022

[8] Steyvers M, Griffiths T. Probabilistic topic models//Landauer T, McNamara D, Dennis S, Kintsch W eds. Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2006

[9] Minka Thomas, Lafferty John. Expectation-propagation for the generative aspect model//Proceedings of the Uncertainty in Artificial Intelligence (UAI). Edmonton, Alberta, Canada, 2002: 352-359

[10] Wang Xiao-Gang, Grimson Eric. Spatial latent dirichlet allocation//Proceedings of the Neural Information Processing Systems (NIPS2007). Vancouver, B.C., Canada, 2007

[11] McCallum A, Corrada-Emmanuel A, Wang X. Topic and role discovery in social networks//Proceedings of the 19th Joint Conference on Artificial Intelligence. Edinburgh, Scotland, 2005: 786-791

[12] Boyd-Graber J, Blei D, Zhu X. A topic model for word sense disambiguation//Proceedings of the Empirical Methods in Natural Language Processing. Prague, 2007

[13] Kristina Toutanova, Mark Johnson. A Bayesian LDA-based model for semi-supervised part-of-speech tagging//Proceedings of the Neural Information Processing Systems (NIPS2007). Vancouver, B.C., Canada, 2007

[14] Newman D J, Block S. Probabilistic topic decomposition of an eighteenth century newspaper. Journal American Society for Information Science and Technology, 2006: 753-767

[15] Wei Xing, Croft Bruce. LDA-based document models for Ad-hoc retrieval//Proceedings of the 29th Annual International SIGIR Conference. Washington, USA, 2006: 178-185

[16] Li Hang, Yamanishi Kenji. Topic analysis using a finite mixture model. Information Processing & Management, 2003, 39(4): 521-541

[17] Croft W B, Cronen-Townsend S, Larvrenko V. Relevance feedback and personalization: A language modeling perspective//Proceedings of the DELOS Workshop: Personalization and Recommender Systems in Digital Libraries. 2001

[18] Kehagias A, Fragkou P, Petridis V. Linear text segmentation using a dynamic programming algorithm//Proceedings of the European Association of Computational Linguistics. Budapest, Hungary, 2003: 171-178

[19] Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation. Machine Learning, 1999, 34: 1-34

[20] Pevzner L, Hearst M. A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics, 2002, 28(1): 19-36

[21] Zhu Jing-Bo, Ye Na, Luo Hai-Tao. Text segmentation model based on multiple discriminant analysis. Journal of Software, 2007, 18(3): 85-94(in Chinese)
(朱靖波, 叶娜, 罗海涛. 基于多元判别分析的文本分割模型. 软件学报, 2007, 18(3): 85-94)

[22] Shi Jing, Dai Guo-Zhong. Text segmentation based on PLSA model. Journal of Computer Research and Development, 2007, 44(2): 242-248(in Chinese)
(石晶, 戴国忠. 基于 PLSA 模型的文本分割. 计算机研究与发展, 2007, 44(2): 242-248)

[23] Kauchak D, Chen F R. Feature-based segmentation of narrative documents//Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. Ann Arbor, MI, USA, 2005: 32-39

附 录.

本文使用的概率模型为

$w_i|\ z_i, \varphi^{(z_i)} \sim Discrete(\varphi^{(z_i)})$, $\varphi^{(z_i)} \sim Dirichlet(\chi)$,

$z_i|\ \psi^{(d_i)} \sim Discrete(\psi^{(d_i)})$, $\psi^{(d_i)} \sim Dirichlet(\alpha)$.

根据 Bayes 公式, z_i的条件后验概率由下式给出:

$$P(z_i=j|\ z_{-i}, w) \propto P(w_i|\ z_i=j, z_{-i}, w_{-i})P(z_i=j|\ z_{-i}) \tag{1}$$

由于参数φ仅出现在右侧表达式的第 1 项, 对其进行积分:

$$P(w_i|\ z_i=j, z_{-i}, w_{-i}) = \int P(w_i|\ z_i=j, \varphi^{(j)})P(\varphi^{(j)}|\ z_{-i}, w_{-i})d\varphi^{(j)} \tag{2}$$

其中$\varphi^{(j)}$是与主题 j 联系的关于词汇的多项分布, 由 Bayes 公式得

$$P(\varphi^{(j)}|\ z_{-i}, w_{-i}) \propto P(w_{-i}|\ \varphi^{(j)}, z_{-i})P(\varphi^{(j)}) \tag{3}$$

因为 $P(\varphi^{(j)}) \sim Dirichlet(\chi)$, 并且是多项分布 $P(w_{-i}|\ \varphi^{(j)}, z_{-i})$的自然共轭先验分布, 于是后验概率 $P(w_{-i}|\ \varphi^{(j)}, z_{-i}) \sim Dirichlet(\chi + n^{(w)}_{-i,j})$. 既然式(2)的最右项仅仅包括$\varphi^{(j)}$, 完成积分得

$$P(w_i|\ z_i=j, z_{-i}, w_{-i}) = \frac{n^{(w)}_{-i,j}+\chi}{n^{(\cdot)}_{-i,j}+W\chi} \tag{4}$$

与此类似, 对式(1)右侧的第 2 项进行积分:

$$P(z_i=j|\ z_{-i}) = \int P(z_i=j|\ \psi^{(d_i)})P(\theta^{(d_i)}|\ z_{-i})\,\mathrm{d}\,\psi^{(d_i)} = \frac{n^{(d_i)}_{-i,j}+\alpha}{n^{(d_i)}_{-i,\cdot}+T\alpha} \tag{5}$$

结合式(4)、(5)有

$$P(z_{-i}=j|\ z_{-i}, w) \propto \frac{n^{(w)}_{-i,j}+\chi}{n^{(\cdot)}_{-i,j}+W\chi} \cdot \frac{n^{(d_i)}_{-i,j}+\alpha}{n^{(d_i)}_{-i,\cdot}+T\alpha} \tag{6}$$

式(6)是非标准化的概率分布, 真正计算时需要除以所有词汇主题分配的概率之和.

SHI Jing, born in 1970, Ph. D.. Her current research interests include natural language understanding and information processing.

HU Ming, born in 1963, Ph. D.. Research interest includes artificial intelligence.

SHI Xin, born in 1975, associate professor. Research interest includes computer application.

DAI Guo-Zhong, born in 1944, research professor, Chief engineer. His research interests include software engineering and computer graphics technology. He is engaged in the development of graphic standard software and user interface.

Background

The research is supported by the National Natural Science Foundation of China. Existing work of text segmentation falls into one of two categories, lexical cohesion methods and multi-source methods. The former proposes that text segments with similar vocabulary and likely to be part of a coherent topic segment. Implementations of this idea use word stem repetition, context vectors, entity repetition, semantic similarity, word distance model and word frequency model to detect cohesion. Approaches for finding the topic boundaries include sliding window, lexical chains, dynamic programming, agglomerative clustering and divisive clustering. Multi-source methods utilize lexical cohesion metrics, cur phrases, prosodic features, ellipsis, anaphora, syntactic features, and language models to detect topic boundaries. Features are combined using decision trees, probabilistic models and maximum entropy models. Text segmentation targets on getting the structure of a text, and therefore is very useful in information retrieval, summarization, text understanding, anaphora resolution, language modeling and text navigation. Most researches aim at the applications in information retrieval. Although many researches have done on text segmentation, few trials are based on LDA model. The work in this paper introduces an approach to segment a document with word distribution computed using LDA model.